- ASSIGNMENT -

Decision Tree

Goal: In this assignment, we will focus on healthcare. This data set is made available by the Center for Clinical and Translational Research, Virginia Commonwealth University. It contains data about 10 years of clinical care at 130 US Hospitals. Each row represents a single patient. The columns include the characteristics of deidentified diabetes patients. This is a binary classification task: predict whether a diabetes patient is readmitted to the hospital within 30 days of their discharge (1=Yes, 0=No). This is an important performance metric for hospitals as they try to minimize these types of readmissions.

1) Start SAS Enterprise Miner and open your project called "Assignments" (if you don't have this project, please create it using the instructions provided in the "Getting Started" section).

2) Create a new diagram called "Decision tre assignment".

3) Click on the plus sign next to "Data Sources" and check whether the "**Diabetes**" data set is already imported into the project. If not, import it from the "mydata" library into this project. (Please see the instructions provided in the "Getting Started" section on how to do this.)

4) **Description of the data set**: The description of each variable is provided on the next page. For the purposes of this assignment, we will focus on the “readmitted” variable and predict whether a patient is readmitted to the hospital within 30 days or not (1=Yes, 0=No).

5) Please set the correct role and level of each variable as shown on the next page. To do this, right click on the data source and select "Edit Variables". In the window that opens, set the role and level of each variable by selecting the right role and level from the dropdown menus.

6) Build at least two decision tree models (possibly more) to compare and minimize the misclassification rate.

Write-up:
To complete this assignment, draw upon what you have learned in the associated exercise and create a write-up. Your write-up should include the following sections:

A. Description of the best decision tree model you identified:
   a. Provide the list of all nodes from start to finish you used for the best model (i.e., list the nodes that concern only the best model.)
   b. List the variable(s) you chose to exclude from the analysis. Discuss the reason for excluding the variable(s). Provide a screenshot of all variables' roles and levels.
   c. For each node: provide the settings you used. (Don't list all the settings. List only the settings you changed. If you didn't change any settings, indicate you used the default settings).
   d. Report the "misclassification rate" value of the model (for validation set). Also compare the misclassification rate to the "baseline" value.

B. Description of the second-best decision tree model:
   a. Provide the list of all nodes from start to finish you used for the second-best model (i.e., list the nodes that concern only the second-best model.)
   b. List the variable(s) you chose to exclude from the analysis. Discuss the reason for excluding the variable(s). Provide a screenshot of all variables' roles and levels.
   c. For each node: provide the settings you used. (Don't list all the settings. List only the settings you changed. If you didn't change any settings, indicate you used the default settings).

d. Report the "misclassification rate" value of the model (for validation set). Also compare the misclassification rate to the "baseline" value.

C. (Optional) Discussion of alternative models:
   a. Use one or more alternative models (such as MBR, regression, neural network).
   b. Report the misclassification rates generated by these models. Compare these values to the misclassification rate of the best decision tree model (as well as the baseline). Discuss which model performs the best (and why).
   c. For each alternative model used:
      i. Discuss the settings used for the model (or say “default” if no settings were changed).
      ii. Discuss if these models use the same preceding nodes. If there are new nodes, provide them and their settings.

D. APPENDIX:
   a. The screenshot of the final diagram
   b. The screenshot of the results window of the best model
   c. The screenshot of the results window of the second-best model
   d. (If available) The screenshot of the results window of the alternative model built in Step C.

Strategies for building a good model:

- Perform data cleaning (if needed) (example: filter outliers, transform variables, etc.) (Try not to transform the target variable)
- Use model comparison to evaluate multiple models.

Description of the Variables

Variable:	Description:	Role	Level
A1Cresult	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.	Input	Nominal
admission_source	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	Input	Nominal
admission_type	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	Input	Nominal
age	Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90, 100)	Input	Nominal
change	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	Input	Nominal
diabetesMed	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	Input	Nominal
diag_1	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	Input	Nominal
diag_2	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	Input	Nominal
diag_3	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	Input	Nominal
discharge_disposition	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	Input	Nominal
gender	Values: male, female, and unknown/invalid	Input	Nominal

insulin	Medication. “up” if the dosage was increased, “down” if it was decreased, “steady” if it did not change, and “no” if the drug was not prescribed	Input	Nominal
max_glu_serum	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured	Input	Nominal
medical_specialty	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon	Input	Nominal
num_lab_procedures	Number of lab tests performed during the encounter	Input	Interval
num_medications	Number of distinct generic names administered during the encounter	Input	Interval
num_procedures	Number of procedures (other than lab tests) performed during the encounter	Input	Interval
number_diagnoses	Number of diagnoses entered to the system	Input	Interval
number_emergency	Number of emergency visits of the patient in the year preceding the encounter	Input	Interval
number_inpatient	Number of inpatient visits of the patient in the year preceding the encounter	Input	Interval
number_outpatient	Number of outpatient visits of the patient in the year preceding the encounter	Input	Interval
patient_id	Unique identifier of a patient	ID	Nominal
payer_code	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay	Input	Nominal
race	Values: Caucasian, Asian, African American, Hispanic, and other	Input	Nominal
readmitted	Readmission. 1 if the patient was readmitted in less than 30 days, 0 otherwise.	Target	Binary
time_in_hospital	Integer number of days between admission and discharge	Input	Interval